

TP2 : Word Embeddings (Plongement lexicaux)

Chaimae EL HOUJJAJI

31 janvier 2022

1 Instructions et Context

1.1 Presentation des corpus

Nous avons 2 corpus sur deux thèmes différents : un sur le domaine médical et un sur le domaine de la presse. Les deux fichiers sont :

- `QUAERO_FrenchMed_traindev.ospl`. Il s'agit d'un corpus en français dans le domaine médical qui est composé de **3091 lignes et 51896 mots** (donnée récupéré en utilisant la commande `wc`)
- `QUAERO_FrenchPress_traindev.ospl`. Il s'agit d'un corpus en français dans le domaine de la press qui est composé de **38548 lignes et 1252121 mots**

Ainsi le corpus sur la presse contient environ **24 fois plus de mots** que le corpus médical et environ 12 fois plus de lignes. Ainsi le corpus médical est un corpus de petite taille par rapport au corpus non médical.

1.2 Embeddings de mots

En 2013, une nouvelle méthode de vectorisation de texte appelée words embeddings est apparu dans le domaine du NLP. Ces nouvelles techniques de vectorisation de texte se distinguent des précédentes méthodes de vectorisation de mots. En effet, elles ont réussi à conserver la similarité sémantique entre les mots, ce qui signifie que ces vecteurs peuvent reconnaître le sens d'un mot et déterminer sa similarité avec d'autres. C'est ce que nous allons testé dans ce projet et nous allons également comparer ces approches.

1.3 Packages utilisés

Afin de faire ces plongements lexicaux (word embeddings), différentes méthodes sont apparues progressivement :

- 2013 : `Word2VC` par Thomas Mikolov chez Google.
- 2014 : `GloVe` par Jeffrey Pennington à Stanford.
- 2016 : `fastText` par Piotr Bojanowski chez Facebook.

Dans le cadre de notre projet, nous étudierons les approches `Word2VC` et `fastText`. Ces deux approches propose deux scénarios différents pour la formation du réseau de neurones :

- **CBOW** ou "continuous bag-of-words" prédit le mot en fonction d'un certain contexte . Cette méthode fonctionne bien pour les mots courants, mais moins bien pour les mots rares.
- **Skip-Gram** prédit le contexte en fonction du mot. Elle fonctionne bien avec une petite quantité de données d'entraînement et représente correctement les mots ou les phrases rares.

Nous allons donc utiliser les approches `word2vec` (Cbow, skipgram) et `fasttext` (Cbow).

1.4 Objectif du projet

L'objectif de ce projet est de pouvoir comparer différentes approches de "word embeddings" appliqués sur des corpus de taille différentes et sur des domaines différents. Pour cela, nous allons dans un premier temps apprendre les embeddings de mots pour chaque approche. Cela nous donnera au final 6 embeddings.

Puis nous allons étudier la similarité sémantique, c'est à dire trouver les 10 mots les plus proches d'un mot donné en s'appuyant sur le calcul de similarité cosinus. Ces mots sont : **patient**, **traitement**, **maladie** et **jaune**. On obtient ainsi pour chacun de

Word2Vec SkipGram semble donner plus de mots en commun.

Nous avons donc essayé de visualiser cela graphiquement. Pour cela nous sommes donc passé d'une représentation des vecteurs avec 100 dimensions à une représentation en 2 dimensions grâce à une TSNE. (Une autre approche possible aurait été de faire une ACP). Le script pour réaliser cela est [Viz_med_wordembeddings.py](#).

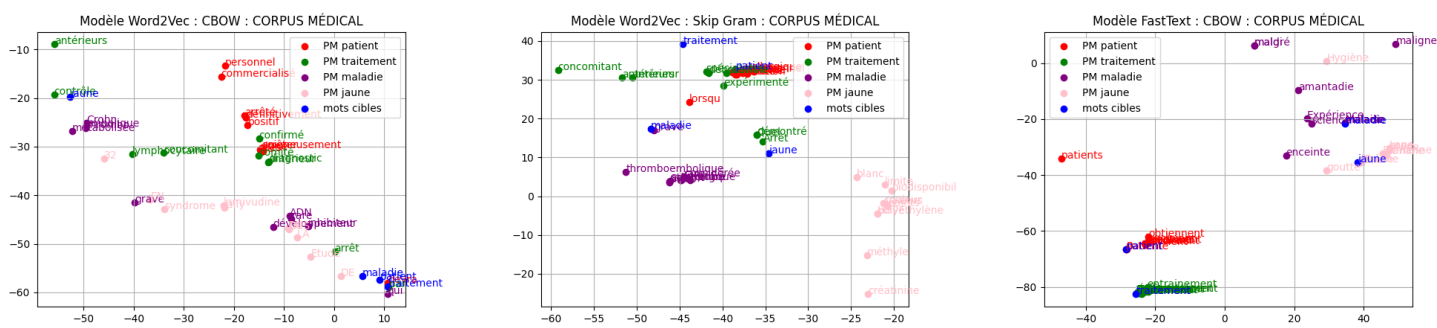


FIGURE 2 – Représentation graphique des mots les plus proches pour 3 approches sur un corpus médical.

D'après la Figure 2, on remarque que dans l'approche Word2Vec CBOW, le mot jaune se distingue bien des mots maladie, traitement et patient qui sont quant à eux regroupés. Toutefois on arrive pas à bien distinguer les clusters associés aux mots cibles et à leurs mots les plus proches.

Ces clusters sont toutefois bien plus visibles dans la représentation avec le modèle Word2Vec Skipgram et avec FastText.

2.2.2 Corpus Non-Médical (de la presse)

Nous allons ensuite comparer les 10 mots les plus proches obtenues par comparaison des 3 approches (Word2vec CBOW, Word2vec SkipGram et FastText CBOW) sur le corpus non médical qui est plus long mais où les mots cibles (patients, maladie, traitement, jaune) font partie d'un vocabulaire du domaine différent : le domaine médical.

patient				traitement				maladie				jaune			
w2v CBOW		w2v Sk	ft CBOW	w2v CBOW		w2v Sk	ft CBOW	w2v CBOW		w2v Sk	ft CBOW	w2v CBOW		w2v Sk	ft CBOW
malaise	0,888	aise	0,956	patientent	0,945	poumon	0,933	alimentaire	0,897	retraitement	0,918	résolution	0,866	assurance	0,870
consensus	0,883	complément	0,953	impatient	0,940	coût	0,925	mépris	0,890	suitement	0,888	responsabilité	0,864	susceptible	0,782
souffle	0,881	clown	0,951	détient	0,926	contenu	0,922	financement	0,887	recrutement	0,884	puissance	0,860	bénéfice	0,761
emprunt	0,877	insupportable	0,951	impatientent	0,909	lycée	0,918	destiné	0,884	bêtement	0,873	perspective	0,851	potentiel	0,761
diplôme	0,874	irréaliste	0,949	renient	0,903	cancer	0,917	potentiel	0,881	doctement	0,872	constitution	0,849	garantie	0,759
découpé	0,873	circonstance	0,949	obtient	0,898	viol	0,916	coût	0,880	vêtement	0,869	compagne	0,843	arme	0,757
revendiquant	0,865	contradictoire	0,948	initient	0,896	poids	0,912	informatique	0,880	abruptement	0,862	discipline	0,843	frappe	0,757
excellent	0,864	journalistique	0,948	dénient	0,893	pain	0,911	outil	0,880	dépècement	0,860	base	0,841	physique	0,755
maître	0,861	motivation	0,947	abstient	0,892	degré	0,910	fiscale	0,877	gratuitement	0,855	réunification	0,840	incapacité	0,753
Prizren	0,861	mollah	0,947	prient	0,892	jazz	0,910	collectif	0,874	tristement	0,855	caisse	0,835	favorise	0,753
														malnutrition	0,632

FIGURE 3 – 10 Mots les plus proches de 4 mots : patient, traitement, maladie et jaune par 3 approches différentes pour le corpus non médical.

D'après la Figure 3, on remarque que les mots les plus proches des 4 mots cibles, peu importe l'approche n'ont que très peu de lien entre eux. On a donc voulu tracer cela dans la Figure 4 graphiquement pour voir ce que cela donne visuellement. (Le script utilisé pour cette visualisation graphique est : [Viz_press_wordembeddings.py](#)).

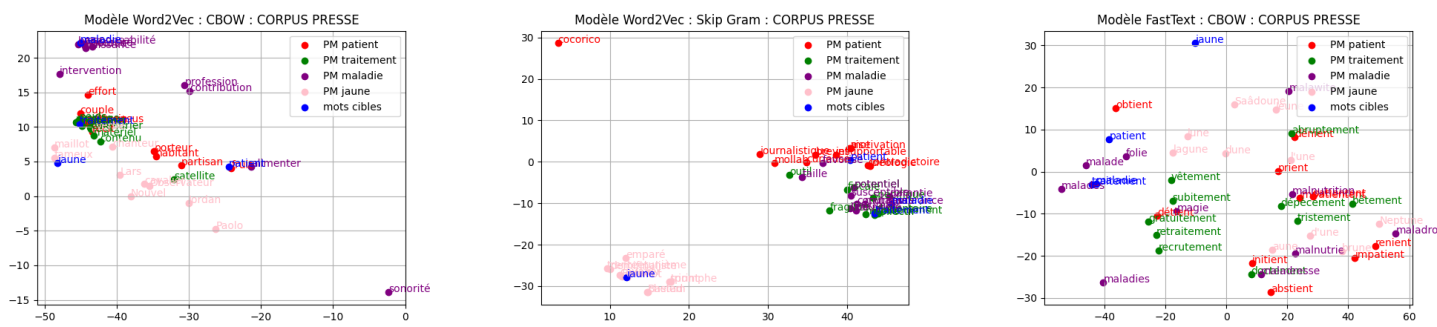


FIGURE 4 – Représentation graphique des mots les plus proches pour 3 approches sur un corpus non médical.

On remarque graphiquement que pour l’approche FastTex, les clusters de mots les plus proches ne se distinguent pas bien. Pour l’approche Word2Vec CBOW on voit une légère répartition en diagonale des clusters. Pour Word2Vec Skipgram par contre la répartition des clusters est plus nette notamment pour le mot ”jaune” qui se distingue des 3 autres mots.

Ainsi on remarque que l’approche Word2Vec SkipGram a tendance a donner de meilleure résultats pour notre jeu de donnée que l’approche Word2Vec CBOW et que l’approche FastText.

2.3 Comparaison des embeddings (même approche) entraînés sur deux corpus différents (médical et non médical)

Nous souhaitons maintenant comparer les 3 approches pour les deux jeux de données : médical et non médical (presse)

2.3.1 Word2Vec avec CBOW

Nous récapitulons les résultats obtenus des 10 mots les plus proches pour 4 mots cibles sur les corpus médicaux et non médicaux avec l’approche Word2Vec CBOW sur le Tableau suivant (Figure 5) et le graphique suivant (Figure 6).

w2v CBOW															
patient				traitement				maladie				jaune			
MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL	
soigneusement	0,986	malaise	0,888	par	0,968	poumon	0,933	directement	0,986	résolution	0,866	LA	0,997	Michael	0,972
lorsqu	0,982	consensus	0,883	exemple	0,968	coût	0,925	ADN	0,985	responsabilité	0,864	DU	0,997	Joey	0,965
allergique	0,979	souffle	0,881	positif	0,958	contenu	0,922	la	0,985	puissance	0,860	Etude	0,997	Spadea	0,964
signe	0,979	emprunt	0,877	arrêt	0,956	lycée	0,918	rare	0,985	perspective	0,851	EN	0,997	maillot	0,962
confirmé	0,977	diplôme	0,874	diagnostic	0,956	cancer	0,917	Crohn	0,982	constitution	0,849	DE	0,997	Kahn	0,962
devra	0,977	découpé	0,873	confirmé	0,954	viol	0,916	embolique	0,981	compagne	0,843	DES	0,996	alias	0,961
particulièrement	0,974	revendiquant	0,865	comité	0,953	poids	0,912	groupe	0,981	discipline	0,843	syndrome	0,996	Jennifer	0,959
il	0,974	excellent	0,864	antérieur	0,951	pain	0,911	prudence	0,980	base	0,841	lamivudine	0,996	Edelmann	0,958
agir	0,972	maître	0,861	affections	0,950	degré	0,910	fin	0,980	réunification	0,840	Lilly	0,996	Jim	0,958
aptitude	0,971	Prizren	0,861	TYSABRI	0,947	jazz	0,910	levure	0,980	caisse	0,835	32	0,995	Perez	0,957

FIGURE 5 – 10 mots les plus proches de 4 mots cibles par une approche Word2VC CBOW

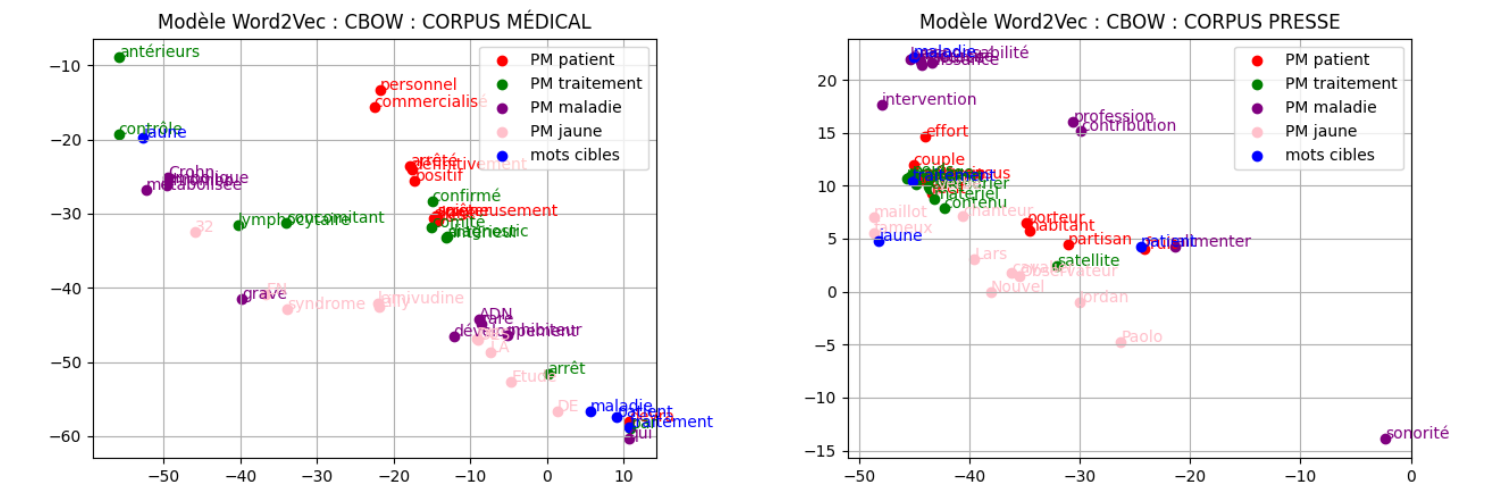


FIGURE 6 – Représentation graphique - Word2Vec CBOW - Medical VS non Médical

2.3.2 Word2Vec avec Skip-Gram

Nous récapitulons les résultats obtenus des 10 mots les plus proches pour 4 mots cibles sur les corpus médicaux et non médicaux avec l'approche Word2Vec Skip-gram (Figure 7) et le graphique suivant (Figure 8).

w2v Skipgram											
patient				traitement				maladie			
MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL	
souffre	0,847	aise	0,956	concomitant	0,790	alimentaire	0,897	thrombo	0,902	assurance	0,870
aptitude	0,847	complément	0,953	début	0,788	mépris	0,890	grave	0,902	susceptible	0,782
modifier	0,842	clown	0,951	Arrêt	0,784	financement	0,887	atteint	0,899	bénéficie	0,761
devrait	0,829	insupportable	0,951	mois	0,783	destiné	0,884	thromboembolique	0,891	potentiel	0,761
conscient	0,828	irréaliste	0,949	expérimenté	0,782	potentiel	0,881	Crohn	0,888	garantie	0,759
personnel	0,825	circonstance	0,949	Traitement	0,781	coût	0,880	embolique	0,882	arme	0,757
présente	0,824	contradictoire	0,948	antérieurs	0,775	informatique	0,880	considérée	0,882	frappe	0,757
analgésique	0,822	journalistique	0,948	résultat	0,770	outil	0,880	malgré	0,872	physique	0,755
allergique	0,822	motivation	0,947	définitif	0,769	fiscale	0,877	rare	0,870	incapacité	0,753
remarquer	0,820	mollah	0,947	antérieur	0,769	collectif	0,874	moelle	0,853	favorise	0,753

FIGURE 7 – 10 mots les plus proches de 4 mots cibles par une approche Word2VC Skip Gram

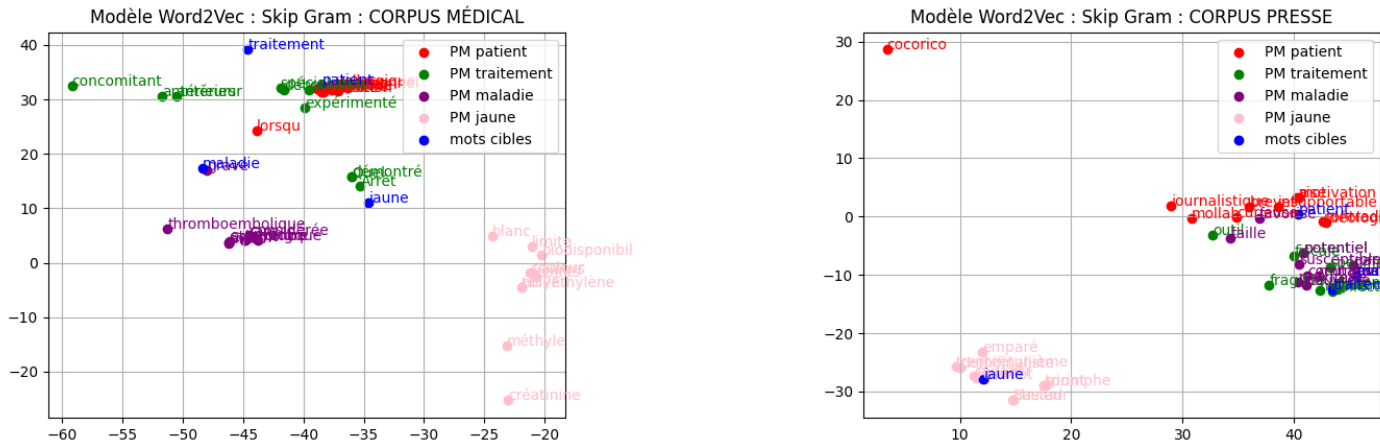


FIGURE 8 – Représentation graphique - Word2Vec CBOw - Medical VS non Médical

2.3.3 FastText avec CBOw

Nous récapitulons les résultats obtenus des 10 mots les plus proches pour 4 mots cibles sur les corpus médicaux et non médicaux avec l'approche FastText CBOw (Figure 9) et le graphique suivant (Figure 10).

FastText CBOw											
patient				traitement				maladie			
MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL		MEDICAL		NON MEDICAL	
Patient	0,987	patientent	0,945	Traitement	0,987	retraitement	0,918	malade	0,980	malade	0,787
patiente	0,961	impatient	0,940	Taaitement	0,978	subitement	0,888	Maladie	0,972	maladies	0,767
parvient	0,948	détient	0,926	traitement	0,973	recrutement	0,884	maldi	0,951	malawite	0,700
appartient	0,945	impatientent	0,909	Allaitement	0,973	bêtement	0,873	malgré	0,919	malnutrie	0,693
maintient	0,945	renient	0,903	allaitement	0,968	doctement	0,872	malaise	0,918	maladresse	0,672
recevaient	0,928	obtient	0,898	étroitement	0,967	vêtement	0,869	amantadie	0,885	folie	0,661
patients	0,918	initient	0,896	évitement	0,965	abruptement	0,862	maligne	0,877	magie	0,654
gradient	0,915	dénient	0,893	entraînement	0,951	dépècement	0,860	malin	0,877	malades	0,643
excipient	0,915	abstient	0,892	département	0,949	gratuitement	0,855	intolérance	0,858	maladroit	0,635
passant	0,915	prient	0,892	recrutement	0,948	tristement	0,855	Parkinson	0,852	malnutrition	0,632

FIGURE 9 – 10 mots les plus proches de 4 mots cibles par une approche FastText CBOw

